

EXPRESS MAIL NO. 9505 9505 L029404559US

PATENT  
DOCKET NO. 00-4037

United States Patent Application

of

Joshua Bers, Patrick M. Peterson and John A. Golden

for

## SYSTEM AND METHOD FOR MAXIMUM BENEFIT ROUTING

[illegible]

**SYSTEM AND METHOD FOR MAXIMUM BENEFIT ROUTING****BACKGROUND OF THE INVENTION****Field of the Invention**

5 [0001] This invention relates generally to a system and method for maximum benefit routing of telephone inquiries. More particularly, the invention relates to a system and method of routing of telephone calls based on identified caller goals and the cost and/or benefit of routing the call to a customer care call center best equipped to respond to the needs of the caller.

**Description of the Related Art**

10 [0002] Telephonic inquiries have become commonplace, and commercial entities that receive many inquiries use automated systems to answer questions or route the call to an appropriate operator. Commercial entities include telephone companies, cable TV providers, long distance telephone providers, public utilities and credit card companies. In many instances the automated phone system presents the calling party with a self-  
15 routing option via touch-tone menus, for example, by asking the caller to select a choice by depressing numbers on the caller's telephone keypad. Sometimes the caller may be confronted with the choice of depressing a number on the keypad or saying the number. The routing system recognizes the tone associated with the depressed number on the keypad, or the vocalized number, and routes the call according to preprogrammed  
20 instructions.

[0003] These simple routing systems work well when the number of routing choices is below 5. As the number of choices increases, multi-tiered menus become necessary. However, callers generally do not like multi-tiered menus, and may be confused by the choices. Accordingly, the percentage of successful routings decreases as the number of  
25 routing choices increases. That is, when faced with a multi-tiered menu system, users frequently have trouble mapping their concerns to the menu choices presented. Thus, callers are less likely to be successfully routed, and in the majority of cases, the call will

be either terminated without the caller having received the desired benefit from placing the call or the call will be defaulted to an operator for disposition.

[0004] Additionally, self-routing via touch-tone menus can be confusing and dangerous for mobile phone users. The use of mobile phones requires that users focus  
5 their visual attention on the handset in order to navigate the phone menu, and not on the driving environment.

[0005] Other techniques include using a topic spotter to route the caller based on spoken responses to a general prompt, where a caller is presented with a simple greeting such as "How may I help you?" The caller responds with a natural speech statement of  
10 the caller's objective, and the system attempts to classify the caller's request into one of a number of predefined objective routings, or to an operator if the request did not fit one of the predefined routings or if the system was unable to understand the caller's request. These systems have the limitation of only taking into account caller preference, and force fit a caller's preference into one of the predefined routings.

[0006] These systems do not take into account the cost or benefits to a call center in maximizing the correct routing of a call. When a call is correctly routed, for example, to the best person qualified to answer the query, or to the call center best equipped to respond to the query, the overall cost to respond to the caller decreases while simultaneously increasing caller satisfaction. Accordingly, there is a need for a maximum  
15 benefit routing system and method that allows for the separation of the caller's goals from those of the call center.  
20

#### SUMMARY OF THE INVENTION

[0007] The present invention provides a maximum benefit call routing system that  
25 allows for the separation of a caller's goals from the goals of a call center. It also provides a maximum benefit call routing system that allows for the separation of a caller's goals from the goals of a call center to route the call in such a manner as to respond to the caller's query. Moreover, the invention allows for the separation of a caller's goals from the goals of a call center to route the call in such a manner as to

respond to the caller's query while minimizing the caller's wait time and maximizing the efficient use of call center resources.

[0008] The present invention relates to an automated call routing system that routes a telephone call by responding to a routing objective of a calling party. A speech  
5 recognizer determines at least one phrase from a speech utterance made by the calling party and outputs a digital phrase. A topic identifier receives the digital phrase and converts it to at least one of a word stem and a word class and generates a topic output. A maximum benefit router receives the topic output and determines where to route the telephone call in order to optimize at least one predetermined parameter.

10 [0009] The invention provides a maximum benefit routing system that facilitates making changes to the structure of a call center while allowing customer's goals to evolve independently. It also provides a maximum benefit routing system that determines the best destination for a call given the goals of both a caller and a call center.

[0010] The present invention is directed to a maximum benefit routing system that  
15 routes telephone calls based on the caller's goals and/or the benefit of routing callers to a customer care center most appropriate for retrieving a valid answer for the caller. In general, the cost or benefit is based on the fastest and least expensive way to answer a query posed by a caller. Using a probabilistic model of the caller's goals or call-topics based on a response to a top-level prompt, and a set of functions associating a utility or  
20 benefit with routing those call-topics to destinations within the center, the utility or benefit is measured according to a measurable criteria such as agent time saved. The invention selects the destination for each call that will have the maximum expected benefit. Stated simply, for example, when a caller has a question about billing, the call is  
25 best routed to a person who has special knowledge about billing, and most likely can answer the billing question in the shortest amount of time. If the billing question was routed to a person having special knowledge about, for example, installation, it most likely would take more time to answer a billing question, and the answer might not be as accurate. In this manner, questions about billing are routed to a person best equipped to

answer the question according to the measured criteria, thereby freeing up a specialist that can answer installation questions from another call inquiring about installation.

[0011] Maximum benefit routing allows for the separation of the caller's goals from those of the call center. This distinction facilitates making changes to the structure of the call center, for example, new agent queues, interactive voice response functions (IVR), and the like, while allowing customer goals to evolve independently. When making a routing decision, the system determines the best destination given the goals of the caller and the call center.

[0012] The approach taken with the inventive system is derived from the decision-theory of minimizing conditional risk described in Pattern Classification and Scene Analysis by Duda and Hart. Bayes decision theory, which is a fundamental statistical approach to the problem of pattern classification, is at least partially interpreted. In decision-theoretic terms, an expected loss is called a risk, and an expected loss associated with the taking of an action is known as a conditional risk. A decision rule is a function of the action taken with a particular observation, which in turn suggests an action to take for every possible observation. To reduce randomness and increase specificity, it can be assumed that for every observation  $x$  the action based on the observation assumes one particular action, resulting in the overall risk being the expected loss associated with a given decision rule. When the goal is to make the risk as small as possible, the action to take for the observation can be determined based on these predetermined decision rules. The resulting minimum overall risk is called the Bayes risk and is the best performance that can be achieved. In the system of this invention, the problem to be solved is expressed in terms of maximizing a benefit.

[0013] In another embodiment, the call-center can map call-topics into the routing destinations by defining a benefit matrix. For example, the center may define 50 distinct call-topics but have only 7 routing destinations. One advantage of this approach is that it facilitates switching call-centers or domains without the need to re-label data or retrain the probabilistic topic spotter.

[0014] In another embodiment, the benefit matrix can associate a high benefit to calls correctly routed to specific destinations while assigning high costs to calls incorrectly routed.

[0015] These and other features and advantages of this invention are described in or  
5 are apparent from the following detailed description of the preferred embodiments.

#### BRIEF DESCRIPTION OF THE DRAWINGS

[0016] The preferred embodiments of the present invention will be described in detail, with reference to the following figures, wherein:

10 [0017] Fig. 1 shows a dialogue processing system in which maximum benefit routing of the present invention can be integrated;

[0018] Figs. 2 and 3 show a topic ID flowchart in accordance with the preferred embodiment of the present invention

[0019] Fig. 4 shows a benefit calculation flowchart in accordance with the preferred  
15 embodiment of the invention;

[0020] Fig. 5 shows a routing decision flowchart in accordance with the preferred embodiment of the invention; and

[0021] Figs. 6-7 show results of experimental data of the present invention.

[0022] Throughout the drawing figures, like reference numerals will be understood to  
20 refer to like parts and components.

#### DESCRIPTION OF THE PREFERRED EMBODIMENT

[0023] In general, the present invention relates to the routing of telephone calls based on computer interpretation of human dialogue and the benefit achieved to both the caller  
25 and the called party when the call is properly routed. The invention relates to the apparatus and methods for operating a computer that processes the dialogue, either digital or analog, to generate another signal that is used to determine the maximum benefit routing, and to route the call based on maximum benefit.

**Figure 1**

[0024] Fig. 1 shows a dialogue processing system 10 in which maximum benefit routing of the present invention can be integrated. Dialogue processing system 10 receives speech utterances 16 when a person 12 transmits to a receiver 18 after a call to a third party using system 10 is connected. It will be understood that speech utterances 16 can be either analog, for example, user 12 talking into a telephone microphone, or they can be digital, for example, typed on a keyboard. Additionally, the speech utterances can be visual, for example, lip reading of a user 12, or the use of hand signals. In this manner, speech utterances 16 can be any dialogue coming from a user 12. Similarly, receiver 18 can be a conventional telephone handset microphone for receiving analog voice and converting the analog voice for electrical transmission, or any other known device, for example a camera, for accepting dialogue from a user 12.

[0025] Output 20 from receiver 18 is input to a speech recognition unit 22 for conversion to a usable digital format that can be input 24 to a dialogue processor 14 for dialogue processing. Dialogue processor 14 has a memory 26, a topic identification program 28 stored in memory 26, and a maximum benefit routing determination program 30 stored in memory 26.

[0026] Upon determination by the maximum benefit routing determination program 30 of the optimum, or best available routing of the call, user 12 is connected with call center 32 most able to deal appropriately with user 12 query.

[0027] Processor 14 is preferably implemented on a general purpose computer. However, processor 14 can also be implemented on a special purpose computer, a programmed microprocessor or microcontroller and peripheral integrated circuit elements, an application specific integrated circuit (ASIC) or other integrated circuit, a digital signal processor, a hard wired electronic or logic circuit such as a discrete element circuit, a programmable logic device such as a programmable logic device (PLD), a programmable logic array (PLA), a field programmable gate array (FPGA) or a programmable array logic device (PAL), or the like. In general, any device, capable of

implementing a finite state machine that is in turn capable of implementing the flow charts shown in FIGS. 2-5, can be used to implement a processor 14.

[0028] Memory 26 is preferably implemented using static or dynamic random access memory (RAM). However, memory 26 can also be implemented using one or more of  
5 static RAM, dynamic RAM, read-only memory (ROM), flash memory, hard disk drive, compact disc read-only memory (CD-ROM) drive, compact disc-readable/writable (CD-RW) drive, Floppy Disk Drive, Network Servers or the like.

[0029] As noted above, memory 26 stores topic identification program 28 and maximum benefit routing determination program 30. Topic identification program 28  
10 and maximum benefit routing determination program 30 may be any program capable of implementing their respective portions of the flow charts shown in FIGS. 2-5. Programs 28, 30 can be written in any language compatible with processor 14.

[0030] Speech recognition unit 22 can be any speech recognizer known in the art. When spoken words are utilized by a user, a speech recognition unit 22 that provides  
15 Spoken Language Understanding (SLU) abilities to applications is preferred. The SLU allows users 12 to express themselves using whatever words they naturally use to have computers intelligently understand their meaning. Speech recognition unit 22 generally provides the necessary SLU components to support speech parsing and conversation management. Thus, each utterance 16 can be parsed into words or phrases that are  
20 indicative of the desire of the user.

[0031] In the preferred embodiment, a statistical language model is used to constrain the recognized word sequences. Many speech recognizers and natural language interfaces are known in the art.

[0032] Maximum benefit routing determination program 30 generates an output 34  
25 that directs user 12 to one of call centers 32. In this manner, a user 12 is connected with a call center 32 based upon a determination of the best routing of the call based upon utterance 16.

[0033] In the preferred embodiment, a call center 32 can be described as having  $L$  routing destinations,  $d_i$ , and  $N$  caller topics,  $t_j$ . Maximum benefit routing determination



program 30 takes as input an  $L \times N$  benefit matrix,  $B$ , and an  $N$ -dimensional topic-likelihood vector,  $\mathbf{t}$ , where

$$\mathbf{t} = P(\text{Topics} \mid \text{Words}).$$

Maximum benefit routing determination program 30 then generates an  $L$ -dimensional vector,  $\mathbf{b}$ , where

$$\mathbf{b} = \text{Benefit}(\text{Destinations} \mid \text{Words}),$$

the expected benefits as output.

[0034] The benefit matrix,  $B$ , is generated by a person familiar with the costs of handling calls in call center 32. The rows of  $B$  represent the destinations for calls, while the columns represent the topics that a caller inquiry may contain. The entries of  $B$  define the benefit in seconds of agent time saved by routing to destination  $d_i$  when the caller's topic is  $t_j$ , or  $\text{benefit}(d_i \mid t_j) = b_{ij}$ .

[0035] The caller's topics could be numerous and unwieldy, so a probabilistic topic spotter, or topic identification program 28, is used to generate a topic-likelihood vector,  $\mathbf{t}$ . The entries for this vector,  $t_i$ , express the probability that the topic, random variable,  $t$ , of the call is  $t_i$ , given the evidence,  $\Pr\{t = t_i \mid e\}$ . Here, the evidence is the string of words,  $e = \text{Words} = \text{word}_1, \text{word}_2, \dots, \text{word}_n$ , that was recognized by speech recognizer 22 from the caller's response to an open-ended prompt. Maximum benefit routing determination program 30 then calculates an output vector,  $\mathbf{b} = \mathbf{Bt}$ , which contains the expected benefits of routing caller 12 to destinations 32 given their utterance 16, such that

$$\text{Benefit}(d_i \mid \text{Words}) = \epsilon \langle \text{benefit}(d_i \mid t) \mid \text{Words} \rangle = \sum_{j=1}^N \text{benefit}(d_i \mid t_j) \Pr\{t = t_j \mid \text{Words}\}.$$

[0036]

Maximum benefit routing determination program 30 then chooses the destination that maximizes this sum, or  $\arg \max_i (b_i)$ , to make the most beneficial routing choice.

25

### Figures 2-3

[0037] Figs. 2 and 3 show a topic identification flowchart in accordance with the preferred embodiment of the present invention. The first part of the topic identification

takes words output by speech recognizer unit 22 at step S1000 and converts them into word stems or word classes. This is performed by extracting each word from step S1000 in step S1010, and continues to step S1020. In step S1020, it is determined whether the word extracted in step S1010 is a member of a class. If the word in step S1020 is a member of a class, the word is replaced with a class in step S1030, and continues with step S1040. If the word is not a member of a class, the word is stemmed in step S1050, and continues with step S1040.

[0038] The stemming in step S1050 can be performed using any known stemming algorithm. In the disclosed embodiment, Porter stemming is used. The Porter stemming algorithm is a process for removing the commoner morphological and inflectional endings from words in the English language. For example, the words connected, connecting, connection and connections can be stemmed to connect. The Porter stemmer was originally described in M. F. Porter, "An Algorithm For Suffix Stripping," Program (Automated Library and Information Systems) 14(3), 130-137, July 1980, which is incorporated herein by reference. The original stemmer was coded in B Combined Programming Language, and has since been rewritten in many other languages, including ANSI C, Java and Perl.

[0039] Additionally, in step S1050, some words such as numbers and dates are collapsed into their respective classes. For example, one, two, three can be collapsed into the class "number."

[0040] In step S1040, keyword lookup is performed on the word class or stemmed word from steps S1030 and S1050 from a table of keywords. When the keyword lookup is complete in step S1040, the process continues to step S1060. In step S1060, it is determined whether the word class or stemmed word is found in the keyword lookup table. If the word class or stemmed word is found in the keyword lookup table, the process continues to step S1070. In step S1070, 1 is added to the keyword count and the process returns to step S1010 for the next word. If the word class or stemmed word is not found in the keyword lookup table, the process continues to step S1080. In step S1080, 1 is added to the non-keyword count and the process returns to step S1010 for the next

word. When there are no more words to be extracted in step S1010, the process continues to step S1090.

[0041] In this manner, topic identification program 28 looks up each resulting stemmed word or word class in a table of keywords that make up the features of a multinomial model.

[0042] In the multinomial model of topic identification program 28 there are  $M-1$  keywords plus one feature that matches all non-keyword word-stems, making a total of  $M$  features. In this manner, topic identification program 28 creates a feature vector of size  $M$  that contains the number of times each feature was found in the input message S1000.

10 [0043] Generalized linear models (GLMs) or support vector machines can be used in place of multinomial, or Bayesian, models for topic modeling. However, multinomial models are used in the preferred embodiment. In the following steps, Bayesian decision theory is utilized, where *a priori* probabilities reflect prior knowledge of how likely an event will occur based on past data. In this embodiment, prior  $P(t_j)$  is estimated by the  
15 frequency of messages in topic  $t_j$  among all messages in a predetermined set of previously recorded live messages. In the case where the data is unavailable, but the *a priori* probabilities  $P(t_j)$  are known, a decision rule must follow. When ancillary data can be determined, a state conditional probability density function can be determined. When the data is observed, Bayes Rule shows how the observed data or evidence,  $e$ , changes the *a*  
20 *priori* probability  $P(t_j)$  to the *a posteriori* probability  $P(t_j|e)$ . As mentioned above, the evidence is the string of words recognized by speech recognizer 22. For the multinomial model we approximate this word sequence as a sequence of keywords,  $e \approx m = w_1, w_2, \dots, w_n$  such that  $w_i \in \text{Keywords} \forall i$ . The keyword distribution  $P(w | t_j)$  is estimated by counting frequencies of the keywords in the set of previously recorded messages with true  
25 topic  $t_j$ . If messages have more than one true topic, then a message with at least one of its true topics equal to  $t_j$  is used for this distribution.

[0044] In step S1090, a multinomial model is performed on the feature vector, and uses a keyword probability matrix 36 and topic prior probabilities 38. Keyword

probability matrix 36 and topic prior probabilities 38 may be stored in memory 26 or in a separate database.

- [0045] The algorithm used in the preferred multinomial model of step S1090 requires that for each topic  $t_j$  ( $j = 1, \dots, N$ ) and message  $m$ , the log posterior probability,  $\log(P(t_j | m))$  is computed and used as the topic's score. The model is multinomial over keywords  $w_1, \dots, w_M$ :

$$\begin{aligned}\log(P(t_j | m)) &= \log \frac{P(t_j)P(m | t_j)}{P(m)} = \log(P(t_j)) + \log \prod_{i=1}^M P(w_i | t_j)^{n_i} - \log(P(m)) \\ &= \log(P(t_j)) + \sum_{i=1}^M n_i \log(P(w_i | t_j)) - \log(P(m)),\end{aligned}$$

- where  $n_i$  is the number of times keyword  $w_i$  appears in  $m$ , and  
10  $P(m)$ , the probability of the message, is given by

$$P(m) = \sum_{t_j} P(t_j)P(m | t_j) = \sum_{t_j} \exp \left( \log(P(t_j)) + \sum_{w_i} n_i (\log(P(w_i | t_j))) \right).$$

In the actual computation of  $P(m)$  we use a normalization technique to avoid underflow. Because small log-likelihoods will tend to drive the exponential computation towards zero, first we calculate  $t_{\max} = \arg \max_{t_j} (P(m | t_j))$  then we can normalize the above by

- 15  $\log(P(m | t_{\max}))$  by subtracting it within the exponential and then adding it back in after the summation,

$$\begin{aligned}\log(P(m)) &= \\ \log \left( \sum_{t_j} \exp \left( \log(P(t_j)) + \sum_{w_i} n_i (\log(P(w_i | t_j)) - \log(P(m | t_{\max}))) \right) \right) + \log(P(m | t_{\max}))\end{aligned}$$

- [0046] For all likelihoods  $P(m | t_j)$  where  $j = 1, \dots, N$  to be comparable, the models  $P(w | t_j)$  use the same keywords. This requires that the estimation of  $P(w_i | t_j)$  for  $w_i$  with zero counts in  $t_j$ 's training. This is accomplished using, for example, Bell-Witten backoff or other known backoff techniques.
- 20

[0047] Output 40 from the multinomial model of step S1090 is a vector of posterior possibilities for each topic in the model,  $P(t_j | m)$ . This output 40 is a vector of scores for topics where each score represents confidence that the message is related to topic  $P(t|m)$  and size equals  $N$ , which equals the number of topics.

5

**Figure 4**

[0048] The benefit calculation is shown in Fig. 4. The output of the multinomial model is a vector of posterior probabilities 40 for each topic in the model, that is,  $p(t_j | m)$ . Vector of posterior probabilities 40, also known as topic scores 40 is input into a benefit calculation at step S1100. The benefit calculation at step S1100 takes these posterior probabilities 40 and weights them by benefits from a benefit matrix 42 to determine the expected benefit of routing the caller to each destination 32. Output 44 from the expected benefit computation of step S1100 is a benefit sorted vector of destinations with topic and benefit scores.

[0049] Cost-benefit matrix 42 for benefit analysis is a  $(d_1, \dots, d_L) \times (t_1, \dots, t_N)$  matrix  $B_{ij}$  of "benefits," where  $B_{ij}$  is the benefit of routing a message to a destination  $d_i$  if its true topic is  $t_j$ . Given this matrix, the classifier calculates for each message  $m$  and topic  $t_j$

15

$$Benefit\langle d_i | m \rangle = \sum_{j=1}^N (B_{ij} \langle P\langle t_j | m \rangle \rangle)$$

20 This represents the expected benefit of routing the message to destination  $j$ . The classifier outputs the  $L$ -dimensional vector  $\mathbf{b}$  of benefit scores.

[0050] In most cases the destinations and the topics are the same, for example, when there is a specific destination associated with each topic. However, the number of destinations and topics need not match. This allows the flexibility of collapsing many topics to a single destination or adding alternate destinations for which no topic exists in the model. For example, it is sometimes useful to have an alternate destination when no topic is suitable. This is also known as a rejection. Selection of this destination usually

25

means that the system would reprompt the caller for additional information. Table 1 depicts an example of a benefit matrix.

Table 1 Benefit matrix

<i>Destinations (i)</i>	<i>Topics (t<sub>j</sub>)</i>			
	Account	Balance	Billing	Confirm-payments
Account	40	-10	-10	-10
Balance	0	80	-10	40
Billing	0	0	10	0
Confirm-payments	-10	-10	-10	80
Reject	0	0	0	0

- 5 [0051] The benefit matrix of Table 1 shows the topics  $t_j$ , the destinations  $d_i$ , and the benefit in agent-seconds. As discussed above, benefit matrix, B, is generated by a person familiar with the costs of handling calls in call center 32. The rows of B represent the destinations for calls, while the columns represent the topics that a caller inquiry may contain. The entries of B define the benefit in seconds of agent time saved by routing to destination  $d_i$  when the caller's topic is  $t_j$ , or  $\text{benefit}(d_i | t_j) = b_{ij}$ . A rejection destination row is added for those instances when the destination is indeterminate. Rejection values are set to zero or some very small value since there is no benefit in sending a caller to reject, or back to an operator to determine how to answer a caller's query. By way of example, when the probabilities output from topic identification program 28 are
- 10 determined to be:
- 15

$$P_1 = \{\text{account}=0.5, \text{balance}=0.4, \text{billing}=0.05, \text{confirm-payments}=0.05\}$$

- then the expected benefit of routing to Account destinations is the dot product of the probabilities output,  $P_1$ , with the Account row, such that the expected benefit score is
- 20

$$\sum (0.5 \cdot 40, 0.4 \cdot (-10), 0.05 \cdot (-10), 0.05 \cdot (-10)) = 15.$$

[0052] In contrast, the expected benefit of routing to Balance is the dot product of  $P_i$  with the Balance row, such that the expected benefit score is

5

$$\sum (0.5 \cdot 0, 0.4 \cdot 80, 0.05 \cdot (-10), 0.05 \cdot 40) = 33.5.$$

The higher sum identifies the maximum benefit destination, or balance. As discussed above, the destination that maximizes this sum, or the max over  $i$  of  $b_i$ , is the most  
10 beneficial routing choice. This example illustrates a case where the maximum benefit destination, Balance, is not the same as the most probable topic, Account. In this manner, the model determines not just the most probable topic, but the maximum benefit destination that will result in the lowest cost.

#### Figure 5

15 [0053] Fig. 5 shows a flow chart of the routing decision. Input to maximum benefit routing determination program 30 is a sorted vector of destinations ordered by their expected benefit score 44. If the destination is also modeled as a topic in the topic id system, the topic score is also included in the vector. If the destination was not modeled as a topic in the topic id system, then a very small value is used. Maximum benefit  
20 routing determination program 30 determines whether to route the call to the top ranking destination or to reject it if the score falls below a predetermined threshold in step S1110. The confidence threshold is determined by plotting the benefit versus the output of the topic id confidence score. The threshold ranges from 0-100, where 100 represents routing only when there is full confidence in the topic. In the preferred embodiment, the  
25 confidence threshold is set between 50-80 in order to receive the maximum benefit. If the topic score is below the threshold value, the utterance is rejected in step S1120. If the topic score is greater than the threshold, the utterance is routed to the top destination in step S1130.

[0054] Using the example from above, maximum benefit routing determination program 30 would receive a benefit vector,

$$B_d = \{Balance = \{33.5, 0.4\}, Account = \{15, 0.5\}, reject = \{0, 0.01\}, Confirm = \{-5.5, 0.05\}\}.$$

- 5 Note that the ranking is based on the benefit calculation but the scores from the topic id output probabilities are also included. Maximum benefit routing determination program 30 would then determine whether to send the caller to the Balance destination if the system confidence threshold was 0.4, or 40% or less, or otherwise reject the utterance, and perhaps reprompting the caller. Other embodiments of the routing decision could use
- 10 both topic score and destination benefit, e.g., route if topic score is above 40% or if benefit is above 30.

#### Data Analysis and Experimental Results

- 15 [0055] Experiments were performed internally to test the effectiveness of the maximum benefit routing approach described above. Data was captured from a large customer care call center that used a wizard-of-oz system to prompt customers, and their response to a routing inquiry, such as, "How may I direct your call?" were recorded. This data was transcribed and each utterance was assigned a topic label from a set of 10 topics.
- 20 About 5000 utterances were used to train the classification system and another independent set of 1200 utterances were used to create an internal-use only test set that was used to test the dialogue processing system 10.

- [0056] In collaboration with an internal the call center, a cost-benefit matrix was created that estimated the cost or savings to the call center in agent-seconds of routing
- 25 calls of various topic types to each of the destinations. Additionally, the cost of a misroute on customer satisfaction was calculated to be equivalent to 10 seconds of agent time. Note that it is difficult to quantify the impact of a misroute on customer satisfaction, or for that matter customer satisfaction on costs. The heuristic used was that a misroute



costs the call center  $\frac{1}{4}$  of the cost in agent time to determine the caller's request (40 seconds) or 10 seconds.

### **Figures 6-7**

[0057] The classification system was then applied to the test set to see where each  
5 call was routed. By applying the costs from the benefit matrix to the test results, the  
average benefit in agent-seconds per call was determined. The first test used the system  
that routes callers to the most likely topic as determined by the topic identifier. This  
resulted in a determination that 17.5 agent-seconds were saved per call. The cost benefit  
matrix was then applied in conjunction with maximum benefit routing determination  
10 program 30 to obtain a new routing choice for the test set. This yielded 19.1 agent-  
seconds saved or a 9% increase in savings per call. The rejection threshold for the system  
was varied to determine how the benefit would respond. The results from these  
experiments are illustrated in Figures 6 and 7.

[0058] In Fig. 6, it can be seen that as the threshold for accepting the top result from  
15 the system is raised beyond the 50% confidence level, the advantage of using the  
maximum benefit routing diminishes. This indicates that this technique is useful when a  
routing decision must be made, for example, when rejection has a very high cost or is not  
available. This might be the case when agents are not available to handle a general call,  
for example, the rejection queue, or when the cost of waiting for a general agent is  
20 prohibitive, for example, very long hold times. Figure 7 shows the tradeoff between  
benefit and number of calls rejected.

[0059] Running at an operating point optimized for maximum benefit for each system  
yields 19.4 and 19.6 agent-seconds saved per call respectively or only a 1-% improvement  
for the maximum benefit system. This indicates that the topics are well matched to the  
25 business requirements of the call center. If there was a large discrepancy between the two  
systems this might indicate that the topics could be remapped to the destinations in a  
more productive way, or it might mean that the treatment of the calls could be changed  
such that the benefit matrix is more balanced or "diagonal".

5 [0060] The benefit matrix can also be used to model changing costs within the call center. Benefit matrices for different times of the day or parts of the year where costs change can be defined. Then only the routing system's active benefit matrix requires updating for it to have an impact on routing performance. This is a much less costly process than retraining the topic models or defining new topics.

10 [0061] It will be appreciated by those in the art that the maximum benefit that can be gained by optimizing the routing of a user call can also apply to routing electronic mail messages. For example, if a user 12 sends an electronic message to a public utility, topic identification program 28 can be applied, followed by maximum benefit routing determination program 30. With the use of electronic messages, errors introduced as a result of speech recognition 22 are eliminated. Optimized routing of email inquiries will lead to better, faster and cheap replies.

15 [0062] The benefit matrix can also be used to collapse topics together when there are fewer destinations than topics. This is achieved by building the matrix such that each topic column has a benefit of 1 in the row of the destination where it should be collapsed. If the benefit matrix is made up of only such entries this will cause the benefit calculation to produce the sum of the confidence scores for the topics that collapse into each destination. The result is a maximum likelihood score rather than a maximum benefit score for the destinations.

20 [0063] While the invention has been specifically described in connection with certain specific embodiments thereof, it is to be understood that this is by way of illustration and not of limitation, and the scope of the appended claims should be construed as broadly as the prior art will permit.